THE IMPACT OF HANDLING MISSED DATA ON THE GAMMA REGRESSION RESPONSE

By

Dr. Amira Ibrahim. El-Desokey

Lecturer at Higher Future Institute for Specialized Technological Studies

aeldesokey@gmail.com

Scientific Journal for Financial and Commercial Studies and Research (SJFCSR)

Faculty of Commerce – Damietta University

Vol.5, No.1, Part 1., January 2024

APA Citation

El-Desokey, A.I (2024). The impact of handling missed data on the Gamma regression response, ***Scientific Journal for Financial and Commercial Studies and Research***, Faculty of Commerce, Damietta University, 5(1)1, 273-303.

Website: https://cfdj.journals.ekb.eg/

THE IMPACT OF HANDLING MISSED DATA ON THE GAMMA REGRESSION RESPONSE

Dr. Amira Ibrahim. El-Desokey

Abstract

This paper presents a comprehensive comparison of various missing data approaches in gamma regression analysis. The study evaluates the performance of linear trend at point method, mean imputation method, and three multiple imputation methods (KNN, PMM, and EM) in handling missing data at different positions (top, center, and bottom) of the data range. The maximum likelihood estimation technique is employed to estimate the parameters of the gamma regression model. An empirical example is presented to demonstrate the application of these methods in analyzing factors affecting carbon dioxide emission in Egypt. The findings reveal that multiple imputation methods outperform other approaches in terms of accuracy and precision. This study provides valuable insights into how different missing data techniques can be utilized to enhance the accuracy and precision of gamma regression models. The results have important implications for researchers and practitioners who use gamma regression analysis to investigate various phenomena with missed data.

Key Words: Gama Regression, Maximum Likelihood, missing data, linear trend at point method, mean imputation method, K- Nearest Neighbor Method (KNN), Predictive Mean Matching (PMM), Expectation Maximization Imputation (EM).

1. Introduction

The Gamma regression model is a general linear model that is commonly used to model continuous response variables that are non-negative and have a skewed distribution Kleinbaum et al. (2002). Unlike linear regression, which presumes the response variable follows a normal distribution, it allows for the modeling of non-normal distributions where the variance exceeds the average). It is particularly useful when modeling count data, as it can handle the inherent skewness and heterogeneity in the data. Gamma regression is also often used in survival analysis, Van Buuren (2012) use it to model the critical care unit duration of stay as a function of patients' characteristics, using function specific factors, Robins et al. (2000) simulate the intervals between Kawasaki disease hospitalizations. Over the past years, attempts have been made to extend generalized linear models, and in particular Gamma regression, to a right-censored outcome. For example, Sigrist and Stahel (2011) investigate estimation in a censored Gamma regression model with application to loss given default. Rein et al. (2011) used gamma regression modeling to estimate healthcare costs associated with hepatitis C treatment, concluding that it was a suitable method for cost analysis in healthcare.

Cepeda, (2001) explored the use of residuals in gamma regression models, which are often used to model data that exhibit positive skewness and heteroscedasticity. He compared the performance of different types of residuals in gamma regression models and found that the Pearson and deviance residuals provided the best fit for the model. Additionally, they suggested that researchers should consider the use of residual plots to assess the fit of their gamma regression models.

The gamma distribution is an essential distribution in probability theory and statistics. It has several important special cases, including the Chi-squared distribution, and the exponential distribution, the gamma distribution has gained popularity as a tool for data imputation in recent years. In (2021), Zhipeng et al. introduced a gamma-distributed data imputation approach, which estimates missing values using the maximum likelihood approach. The lifetime distribution of aircraft components was also described using a mixture-Gamma distribution model established by Meng & Rubin (1993) in the context of incomplete data. The study forecasted the failure time of such components using the proposed model. These studies highlight the importance of gamma regression in handling missing data in statistical analysis. Dupuy (2020) proposed different methods for addressing the issue of missing censoring indicators in censored Gamma regression models. Specifically, he suggested using regression calibration, multiple imputations, and augmented inverse probability weighted estimates. Simulation analyses confirmed the effectiveness of the proposed strategies with respect to bias and mean squared error. Overall, the study provides valuable insights into how to handle missing data in censored Gamma regression models.

The paper aims to compare the efficiency of various missing methods in gamma regression. The paper compares the linear trend at point method, The structure of the article is as follows: Section 1: Introduction; Section 2, introduced Gamma Regression Model; Maximum Likelihood Estimation is discussed in Section 3, Section 4, Investigated the Methods for Handling Missing Data in Gamma Regression; and Section 5, analyzed our Case Study Using R programming language.

2. Gamma Regression Model

Suppose the random variable $y_i, i = 1,2,\ldots,n$ the Gamma distribution's positively skew response variable, when both the shape parameter λ and the scale parameter α are positive numbers. This leads us to the formula for the pdf of the response variable:

$$f(y_i) = \frac{\lambda^{\alpha} {y_i}^{\alpha-1} e^{-\lambda y_i}}{\Gamma(\alpha)} \boldsymbol{I}_{(0,\infty)}(y_i), \, y > 0 \qquad (1)$$

When the gamma function is denoted by, α, λ > 0, and the indicator function is denoted by $I(.)$ For this set of parameters, we get the following expressions for the mean and standard deviation of, $E(\mathrm{Y}) = \mu/\alpha$ and $\mathrm{V}\,(Y) = \frac{\mu_i^2}{\alpha}$, Re-parameterizing the gamma distribution function (1) as a function of the mean (μ), and shape (α) parameters as yields the following form for the distribution function:

$$f(y_i) = \frac{1}{y\Gamma(\alpha)} \left(\frac{\alpha y}{\mu}\right)^{\alpha} e^{-\alpha y_i/\mu} \boldsymbol{I}_{(0,\infty)}(y_i) \qquad (2)$$

Where, $\mu, \alpha > 0,$ The gamma function is represented by$\Gamma(.)$, while the indicator function is denoted by $I(.)$. Bossio and Cuervo (2015) propose the notation $y_i \sim G(\mu, \alpha)$ to indicate that y follows a gamma distribution with $\mathrm{E}(y_i) = \mu$ and α as a shape parameter.

Take an n-size random sample, denoted $y_i \sim G(\mu_i, \alpha)$, where i ranges from 1 to n. Gamma regression models with a fixed shape parameters use a model with a regression structure to model data where the mean and shape parameter vary between observations. The parameters for the form and mean of the regressions are defined by:

$$g(\mu_i) = \eta_{1i} = x_i'\beta \tag{3}$$

$$h(\alpha_i) = \eta_{2i} = z_i'\gamma \tag{4}$$

where g and h are adequate actual link features link functions, $\beta = (\beta_0, \beta_1, \dots, \beta_q)'$, and $\gamma = (\gamma_0, \gamma_1, \dots, \gamma_k)$ are respectively, the average and the shape regression parameter vectors, x_i, z_j are, respectively, the average and the shape explanatory variables for the $i\ th$ and observation and The $i\ th$ value of the explanatory variables (represented by the vectors , η_{1i}, and η_{2i}) is a linear predictor (represented by the vectorη_i). Here, $g(.) \colon (0, \infty) \mapsto \Re$ is a strictly monotonic, twice differentiated real value function. The logarithm function is one example of a frequently used mean link function in gamma regression. The identity function, $g(\mu_i) = \mu$, and its inverse, $g(\mu_i) = 1/\mu.$, can be written as $g(\mu_i) = \log(\mu)$. The inverse function is the standard link for the mean in generalised linear models.

3. Maximum Likelihood Estimation

Cepeda-Cuervo, et al. (2016), provided a traditional method for regression of gamma models, based on the Fisher scoring technique, in which the mean and shape parameters also adhere to regression structures. And shown that the likelihood function may be expressed as: when the gamma parameterization described by (2) is used.

$$L(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \prod_{i=1}^{n} \frac{1}{\Gamma(\alpha_i)} \left(\frac{\alpha_i}{\mu_i}\right)^{\alpha_i} y_i^{\alpha_i - 1} \exp\left(-\frac{\alpha_i}{\mu_i} y_i\right) \tag{5}$$

$$l(\boldsymbol{\beta}, \gamma) = \sum_{i=1}^{n} \left\{-\log\left[\Gamma(\alpha_i)\right] + \alpha_i \log\left(\frac{\alpha_i y_i}{\mu_i}\right) - \log(y_i) - \left(\frac{\alpha_i}{\mu_i}\right) y_i\right\} \tag{6}$$

Assuming, then, the regression models described by $\mu_i = x_i'\beta$, and $\alpha_i = z_i'\gamma$, the score statistics are given by:

$$\frac{\partial l}{\partial \beta_j} = \sum_{i=1}^{n} -\frac{\alpha_i}{\mu_i}\left(1 - \frac{y_i}{\mu_i}\right) x_{ij}, j = 1, \dots p$$

$$\frac{\partial l}{\partial \gamma_k} = \sum_{i=1}^{n} -\alpha_i \left[\frac{d}{d\alpha_i} \log \Gamma(\alpha_i) - \log\left(\frac{\alpha_i y_i}{\mu_i}\right) - 1 + \frac{y_i}{\mu_i}\right] z_{ik}, k = 1, \dots, r$$

and Formulas for establishing the Hessian matrix are

$$\frac{\partial^2 l}{\partial\beta_k\,\partial\beta_j} = \sum_{i=1}^{n} \frac{\alpha_i}{\mu_i^2}\left(1-\frac{2y_i}{\mu_i}\right) x_{ij}x_{ik}, j,k = 1,\ldots p$$

$$\frac{\partial^2 l}{\partial\gamma_k\,\partial\gamma_j} = \sum_{i=1}^{n} -\alpha_i\left[\frac{d}{d\alpha_i}\log\,\Gamma(\alpha_i) - \log\left(\frac{\alpha_i y_i}{\mu_i}\right) - 1 + \frac{y_i}{\mu_i}\right] z_{ij}z_{ik}$$

$$-\sum_{i=1}^{n}\alpha_i\left[\alpha_i\frac{d^2}{d\alpha_i^2}\Gamma(\alpha_i)-1\right] z_{ij}z_{ik}, j,k = 1,\ldots,r$$

Formula for the Fisher information matrix is given as

$$-E\left(\frac{\partial^2 l}{\partial\beta_k\beta_j}\right) = \sum_{i=1}^{n} \frac{\alpha_i}{\mu_i^2}x_{ij}x_{ik}, j,k = 1,\cdots,p$$

$$-E\left(\frac{\partial^2 l}{\partial\gamma_k\beta_j}\right) = 0, j = 1,\cdots,p, k = 1,\cdots,r$$

$$-E\left(\frac{\partial^2 l}{\partial\beta_k\beta_j}\right) = \sum_{i=1}^{n} \alpha_i^2\left[\frac{d^2}{d\alpha_i^2}\log\,\Gamma(\alpha_i)-\frac{1}{\alpha_i}\right] z_{ij}z_{ik}, j,k = 1,\cdots,r$$

The Fisher information matrix is a block diagonal matrix, with the two diagonal blocks representing the mean and shape regression parameters, respectively. As a result, the maximum likelihood estimators for and are asymptotically independent, proving that and are orthogonal according to Cox and Reid (1987).

By considering the Fisher information matrix, Cepeda (2001) concludes that the Fisher scored info formula can be expressed as follows:

$$\beta^{(k+1)} = \left(X'W_1^{(k)}X\right)^{-1}X'W_1^{(k)}Y \tag{7}$$

$$\gamma^{(k+1)} = \left(Z'W_2^{(k)}Z\right)^{-1}X'W_2^{(k)}\tilde{Y} \tag{8}$$

A diagonal matrix $\boldsymbol{W}_1^{(k)}$ having elements of the form $w_{ii}^{(k)} = (\mu_i^2/\alpha_i)$, and

$$\tilde{y}_i = \eta_{2i} - \frac{1}{\alpha_i}\left[\frac{\partial^2}{\partial\alpha^2}\log\,\Gamma(\alpha_i) - \frac{1}{\alpha_i}\right]^{-1}\left[\frac{\partial}{\partial\alpha_i}\log\,\Gamma(\alpha_i) - \log\left(\frac{\alpha_i y_i}{\mu_i}\right) - 1 + \frac{y_i}{\mu_i}\right].$$

a diagonal matrix $\boldsymbol{W}_2^{(k)}$ having elements of the form $w_{ii}^{(k)} = 1/d_i$

$$d_i = \alpha_i^{-2}\left[\frac{d^2}{d\alpha_i^2}\log\,\Gamma(\alpha_i) - \frac{1}{\alpha_i}\right]^{-1}$$

And suggested an iterative approach to derive maximum likelihood estimates of the model of the regression parameters based on the structure of the Fisher information matrix as

1. Initialize the iteration count to k = 0.
2. Provide the starting values for β and γ.
3. Given the current values of β and γ, use equation (7) to derive$\beta^{(k+1)}$.
4. Given the current values of β and γ., calculate $\gamma^{(k+1)}$ from equation (8).
5. Put k = k + 1 into the counter iteration.
6. Head for 3 until convergence occurs.

The Fisher information matrix is blocking diagonal for other link functions such as like$g(.) = log(.)$ and $h(.) = log(.)$, allowing for the implementation of an alternative iterated method.

4. Methods for Handling Missed Data in Gamma Regression Model

Missed data are a frequently challenge that can affect the quality of the analysis and lead to biased or inefficient estimates. Therefore, handling missing data in gamma regression is essential to obtain accurate and reliable results. There are several methods available for handling missing data in gamma regression. The missing data's reason should be determined before any action is taken. We investigate four different "missingness mechanisms," starting with the most basic and working our way up to the most complex. There are three general missingness mechanisms commonly used in the statistics literature on missing data: first, when the missing data are unrelated to both the observed and the unobserved data, we say that the data are Missing Completely at Random (MCAR). Second, when missed data is unrelated to other missed data but is linked to part of the observed data, we call this "Missing at Random" (MAR). Third, According to Abonazel and Ibrahim (2018), "Missing Not at Random" (MNAR) occurs when missing data is linked to the way in which they were lost.

4.1 Linear Trend at Point

The linear trend at point (LTP) method is a commonly used approach to assign missing data in a dataset. This approach estimates missing values by compitableing a model of linear regression to the observed values and then using the estimated coefficients to predict the missed data. The LTP approach assumes that the data follows a linear trend, and that the missing values lie in this trend. This method can be effective when the missed data is missed completely at random (MCAR) or missed at random (MAR) and can be used in the case of both discrete and continuous values. However, this method may not perform well when the data has non-linear patterns or when there is substantial missing data, Keating & Tripathy, (2018).

4.2 Mean Imputation

Mean imputation (substitution) is the most used estimation method; this technique replaces missing values for missing cases with the variable mean value. An advantage of this method is that by replacing the missing value with an actual value (mean value), this will increase the sample to its original size, which overcomes the issue of the wasted data produced by using deletion methods techniques.

4.3 K- Nearest Neighbor Method (KNM)

K-nearest neighbor (KNN) method: This method selects the K nearest individuals to the one with the missing value and chooses the donor from this subset. The distance between individuals is calculated using one of several methods, including the Mahalanobis distance method and the Euclidean distance method, van Buuren (2012); Little & Rubin (2019).

4. 4 Predictive Mean Matching (PMM)

Predictive Mean Matching (PMM) is a type of multiple imputation technique that has been developed for handling missing data. In the PMM approach, it is presumed that the missing values on a variable for an individual should be similar to the observed values for that variable in individuals who are similar to the one with a missing value. PMM scans the data and for each missing value, it gathers a set of individuals called donor candidates, who are like the individual with the missing value. Then, the observed value of one of these candidates is imputed to replace the missing value, creating a full data set. PMM is a popular imputation method because it is easy to use and generally provides better results than other approaches as average imputation or deletion. Several studies have shown that PMM is a powerful Technique of Imputation that can handle various kinds of data missing, including missed completely at random, missed at random, and missed not at random, De Waal et al. (2011).

4.5 Expectation Maximization Imputation (EM)

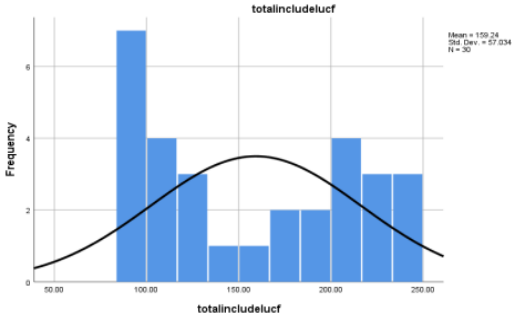
Dempster et al. (1977) created the EM technique, an iterative approach for calculating maximum likelihood estimates in the presence of incomplete longitudinal data. The algorithm has two steps: E-step and M-step, or "expectation" and "maximisation," respectively. In the E-step, we use our best predictions for the model parameters to establish missing values, The M-step then re-estimates the model's parameters using the imputed data. These steps are iterated until convergence is reached. Using the EM technique has become widespread. in various fields such as biostatistics, finance, and natural language processing (NLP) due to its flexibility and ability to handle missing data. Several extensions and modifications of the EM algorithm have also been proposed, including the MCEM (Monte Carlo EM) technique and the SEM (Stochastic EM) approach, McLachlan & Krishnan, (2008); Meng & Rubin, (1993).

5. Case study

We explored the impact of Land Change and Forestry (LUCF) and other factors in Egypt's total Co2 emission in the period 1990- 2019. Electricity and heat, Building, transportation, changing land uses, and forest management are all contributing variables, respectively. We considered that the Total emission include land use change forestry as a dependent variable, while the others variables are the independent variables. We investigate the study in and analyze three examples where data was missing at varying points in the sequence, at the beginning, at the mid-sereis, and at the ending of the series. Applying the methods of handling missed data, we examine the Gamma regression models and determine the one that provides the best fit in each case, We used R programming language for our Study.

5.1. Non- Missing Case

Examining the dependent variable's normality in Fig. (1), we can see that is not distributed normally.



.

The Normality of Data Set

Fig. (1)

We used R language to get the descriptive data; we have provided a summary statistics for four covariates (independent variables) and one dependent variable "Total emissions include lucf". We get the sample size (N), each variable's minimum and maximum values, as well as the mean and standard deviation. For the dependent variable "Total emissions include lucf", the sample size is 30, with a minimum value of 87.34, a maximum value of 249.55, a mean of 159.2407, and a standard deviation of 57.03442.

For the covariate "Electricity and heat", the sample size is also 30, with a minimum value of 25.36, a maximum value of 112.73, a mean of 60.5037, and a standard deviation of 29.60975.

Similarly, for the covariates "Building" and "Transportation", the sample size is 30, with minimum and maximum values of 8.10, 17.03, and 20.51, 40.62, respectively. The means for "Building" and "Transportation" are 12.7530 and 29.6450, respectively, and their standard deviations are 2.86758 and 5.07910, respectively. Finally, for the covariate "land use change and forestry", the sample size is 30, with a minimum value of -0.41, a maximum value of 0.92, a mean of -0.0297, and a standard deviation of 0.47246.

The Gamma Regression Model is given by the following Formula

$$\mu y = c_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$$

5.1.1 Gamma Regression Model with Identity Link Function

This study provides the estimated coefficients from the identity link function in a gamma regression model. The predictor variables included are "Electricity and heat", "Building", "Transportation", and "Land use change and forestry". The "Intercept" term represents the estimated constant term in the model. The p-values for "Electricity and heat", "Building", and "Transportation" are all less than .001, which indicates that these variables are correlated extremely well with the resulting variable. The p-value for "Land use change and forestry" is .085, which is greater than .05 and suggests this variable is not significantly associated in terms of the resultant variable at the .05 the level of significant.

5.1.2 Model of Gamma Regression with Logarithmic Link Function

In this section and with using the R Language programming we noticed that the p-values for "Electricity and heat", "Building", and "Transportation" are all less than .001, which indicates that these variables are correlated significantly with the outcome of the variable. The p-value for "Land use change and forestry" is .145, which is greater than .05, meaning that there is insufficient proof to establish that this independent variable significantly affects the dependent one "Total emissions including LUCF".

Now we use the omnibus test of the model of gamma regression uses both of the identity and log link functions. The omnibus test is used to determine whether the model significantly explains the dependent variable variability. The results presented that the LRχ^2 (model likelihood chi-square) value with the identity link function is 166.283 with 4 degrees of freedom and a p-value of .000. This confirms the identity-link function model. Significantly explains the variability in the dependent variable.

Similarly, the LRχ^2value for the model with the log link function is 115.221 with 4 degrees of freedom and a p-value of .000, indicating that the model with the log link function also significantly explains the variability in the dependent variable.

Overall, these results suggest that both the identity link function and the logarithmic link function are appropriate for modeling the relation between the gamma regression model's dependent variable and its independent variables.

5.1.3 Compare between Models

Table1: goodness of fit

	Identity Link Function	**Log Link Function**
Pesudo R^2	0.995	0.978
Log Likelihood	-79.463	-104.994
Akaike's Information Criterion (AIC)	170.927	221.989
Corrected for a Finite Sample AIC (AICC)	174.579	225.641
Bayesian Information (BIC)	179.334	230.396
Consistent AIC (CAIC)	185.334	236.396

Several indicators of the gamma regression models' goodness of fit are provided in Table 5, with identity and log link functions.

- Pseudo R^2: these measures percentage of variation clarify by the model. The higher valued, indicates a more optimal fit. The pattern, with identity link function has a higher Pseudo R^2 (0.995) than the model with log link function (0.978), indicating a superior fit for the former pattern.
- Log likelihood: This is a measurement of the model's fit to the data, based on the maximum likelihood estimation. The higher valued, the opyimal fit. The model with identity link function has a higher log likelihood (-79.463) than the model with log link function (-104.994), suggesting a better fit for the former model.
- Akaike's Information Criterion (AIC): The likelihood and the number of variables are used to get this measure of the model's quality. The lower valued, the better the fit. The model with identity link function has a lower AIC (170.927) than the model with log link function (221.989), suggesting a better fit for the former model.

- Finite Sample Corrected AIC (AICC): This is a modified version of AIC that corrects for the small sample size. The lower valued, the better the fit. The pattern with identity link function has a lower AICC (174.579) than the model with log link function (225.641), suggesting a better fit for the former model.
- Bayesian Information Criterion (BIC): The likelihood and the number of variables are used to get this measure of the model's quality, the lower valued, the better fit. The model with identity link function has a lower BIC (179.334) than the model with log link function (230.396), suggesting a better fit for the former model.
- Consistent AIC (CAIC): This is a modified version of AIC that penalizes for overfitting and is useful for small sample sizes. The lower valued the better fit. The pattern with identity link function has a lower CAIC (185.334) than the model with log link function (236.396), suggesting a better fit for the former model.

Overall we find that the Identity Link function in the Gamma regression model superiors the Log Link functions.

5.2 Handling Approaches with Missing Data

In this case we are handling missing data in gamma regression in three cases, first when missing data in starting the data using the Identity Link function.

5.2.1 Missing Data in the Top data

5.2.1.1 Linear Trend at Point Method

Table 2: Parameter Estimates

Parameter	B	Std. Error	95% Wald Confidence Interval		Hypothesis Test		
			Lower	Upper	Wald Chi-Square	df	Sig.
(Intercept)	-43.002	32.5728	-106.843	20.840	1.743	1	.187
electricity and heat	.513	.4256	-.321	1.347	1.453	1	.228
Building	9.754	3.8318	2.244	17.264	6.480	1	.011
Transportation	1.670	.7933	.115	3.224	4.429	1	.035
land use change and forestry	12.038	12.5035	-12.469	36.544	.927	1	.336
(Scale)	.009[a]	.0024	.006	.016			

This table provides the results of gamma regression with identity link function after handling missing data with the linear trend at point method, row one of the table displays the intercept coefficient, which represents the response variable's value when all independent variables have a value of zero. The intercept is not statistically significant in this case. (p = .187), indicating that there is no evidence of a non-zero intercept.

The next four rows show the estimated coefficients for each of the independent variables: electricity and heat, building, transportation, and land use change and forestry. The coefficient for electricity and heat is statistically not significant (p > .05), implying that there is no evidence of a linear relation between this variable and the response variable. The coefficients for building and transportation are statistically significant (p = .011 and p = .035, respectively), indicating that there is a positive linear relationship between these variables and the response variable. The coefficient for land use change and forestry is also not statistically significant (p = .316), indicating that no evidence exists of a linear relationship between this variable and the response variable.

The omnibus test for gamma regression is used to determine if the fitted model is significant or not, the results proved that the likelihood ratio chi-square statistic is 77.230, with 4 degrees of freedom and A level of significance lesser than .001. This shows that the fitted pattern is significantly better than the intercept-only pattern, and therefore, after handling missing values, the gamma regression model is an adequate fit for the data.

5.2.1.2 Mean impute Method

Table 3: Estimates of Parameters

Parameter	B	Std. Error	95% Wald Confidence Interval		Hypothesis Test		
			Lower	Upper	Wald Chi-Square	df	Sig.
(Intercept)	-23.113	30.4295	-82.753	36.528	.577	1	.448
electricity and heat	.838	.3728	.107	1.568	5.051	1	.025
Building	8.083	3.6513	.927	15.240	4.901	1	.027
Transportation	1.055	.6313	-.182	2.293	2.794	1	.095
land use change and forestry	7.659	7.6458	-7.326	22.644	1.003	1	.316
(Scale)	139.200^a	35.9412	83.919	230.896			

This table shows the parameter estimates for gamma regression after using mean impute method to handle missing data. The coefficient for electricity and heat (x_1) has a significant positive effect on the dependent variable ($p < .05$), indicating that an increase in electricity and heat production leads to a higher greenhouse gas emission. The coefficient for building (x_2) also has a significant positive effect ($p < .05$), suggesting that building activities contribute significantly to greenhouse gas emissions. The coefficient for transportation (x_3) does not reach statistical significance at the .05 level ($p = .095$), but it shows a positive effect on greenhouse gas emissions. Finally, the coefficient for land use change and forestry (x_r) is not significant ($p = .316$), indicating that this variable may not be a significant predictor of greenhouse gas emissions in this model. The omnibus test is performed using a likelihood proportion chi-square test, with a chi-square value of 77.230 and 4 degrees of freedom. The p-value test is lesser than 0.05, implying that the model is significant.

5.2.1.3 KNN Method

Table 4: Parameters Estimates

Parameter	B	Std. Error	95% Wald Confidence Interval		Hypothesis Test		
			Lower	Upper	Wald Chi-Square	df	Sig.
(Intercept)	-7.476	6.3711	-19.964	5.011	1.377	1	.241
electricity and heat	1.307	.0856	1.139	1.475	233.031	1	.000
Building	4.835	.7455	3.374	6.296	42.067	1	<.001
Transportation	.887	.1578	.578	1.197	31.614	1	<.001
land use change and forestry	5.393	2.5479	.399	10.387	4.481	1	.034
(Scale)	.000[a]	.0001	.000	.001			

This table shows the parameter estimates for a gamma regression model fitting to the imputed dataset using the KNN imputation method. The intercept in this model is -7.476, at a standard error of 6.3711. This indicates the expected value of greenhouse gas emissions when all independent variables are equal to zero. The other independent variables show positive coefficients, indicating that they are positively associated with greenhouse gas emissions. The coefficient for electricity and heat is 1.307, building is 4.835, transportation is 0.887, and land use change and forestry is 5.393. All of the predictor variables have statistically significant coefficients ($p < .05$), indicating that they are associated with carbon dioxide emissions. The intercept coefficient is not statistically significant ($p = .241$), which means that it is not significantly different from zero and may not be necessary to include in the model. Finally, the (Scale) parameter in the model is also shown, which estimates the scale parameter of the gamma distribution. This parameter indicates the degree of data variability that cannot be explained by the independent variables.

Using the omnibus test In this case, the likelihood proportion Chi-square test was used to determine statistical significance of the entire model, and the results show a value of chi-square of 174.395 with 4 degrees of freedom and a p-value of .000, indicating that the model is significant.

5.2.1.4 PMM Method

Table 5: Estimates of Parameters

Parameter	B	Std. Error	95% Wald Confidence Interval		Hypothesis Test		
			Lower	Upper	Wald Chi-Square	df	Sig.
(Intercept)	-6.675	6.7746	-19.954	6.603	.971	1	.324
electricity and heat	1.326	.0911	1.147	1.504	211.689	1	.000
Building	4.724	.7926	3.170	6.277	35.520	1	<.001
Transportation	.869	.1679	.540	1.198	26.774	1	<.001
land use change and forestry	5.225	2.7121	-.091	10.540	3.711	1	.054
(Scale)	.000[a]	.0001	.000	.001			

The results suggest that electricity and heat, building, and transportation have a significance positive effect on the variable that is dependent (not shown in this table), while land use change and forestry has a positive effect but is marginally significant at a 0.05 significance level. The scale parameter estimate suggests that the data have a gamma distribution. The Omnibus test shows that the overall model is significant at a 0.05 significance level. The corresponding omnibus tests are identical to the ones we posted earlier. The omnibus test tests that all regression coefficients in the model are zero as a null hypothesis that, indicating that none of the variables independent are significant predictors of the dependent variable. In this case, the likelihood proportion chi-square statistic is 170.752 with 4 degrees of freedom and a p-value of 0.000, showing significant evidence in contradiction to the null hypothesis and suggesting that at least one variable independent is a significant estimator of the dependent variable.

5.2.1.5 EM Method

Table 6: Parameters Estimates

Parameter	B	Std. Error	95% Wald Confidence Interval		Hypothesis Test		
			Lower	Upper	Wald Chi-Square	df	Sig.
(Intercept)	-8.710	5.9662	-20.404	2.983	2.131	1	.144
electricity and heat	1.279	.0801	1.122	1.436	254.818	1	.000
Building	5.007	.6983	3.638	6.375	51.413	1	<.001
Transportation	.916	.1477	.626	1.205	38.465	1	<.001
land use change and forestry	5.651	2.3821	.982	10.320	5.628	1	.018
(Scale)	.000[a]	8.9835E-5	.000	.001			

The hypothesis test for each parameter tests that the true value of the coefficient is zero as the null hypothesis, indicating that the corresponding independent variable has no impact on the dependent variable. The Wald Chi-Square statistic is utilised for this test. If the p-value associated with the test statistic is less than a chosen significant level (0.05), which implies to the rejection of null hypothesis, indicating that the corresponding independent variable has a statistically significant impact on the dependent variable.

The Omnibus Test in this context is likely referring to a statistical test that evaluates the overall significant of the pattern, often utilizing the likelihood ratio chi-square test. In this case, the output shows that the value of the likelihood ratio chi-square is 178.278, with 4 degrees of freedom, and a significance level of .000. This indicates that the model as a whole is statistically significant, and that there is strong evidence to suggest that the predictors are related to the outcome variable.

5.2.2 Missing Data in the Center data

5.2.2.1 Linear Trend at Point Method

Table 7: Estimates of Parameters

Parameter	B	Std. Error	95% Wald Confidence Interval		Hypothesis Test		
			Lower	Upper	Wald Chi-Square	df	Sig.
(Intercept)	-.085	7.4616	-14.709	14.540	.000	1	.991
electricity and heat	1.442	.1000	1.246	1.638	208.185	1	.000
Building	4.219	.8727	2.509	5.930	23.373	1	<.001
transportatin	.638	.1841	.278	.999	12.026	1	<.001
land use change and forestry	3.250	2.9901	-2.610	9.111	1.182	1	.277
(Scale)	.001[a]	.0001	.000	.001			

Based on the given table, we can interpret the significance of each variable as follows:

Intercept: The p-value of the intercept is 0.991, which is superior to the commonly used significant level of 0.05. Therefore, we fail to reject the null hypothesis that the intercept is equal to zero. This suggests that the model does not need a constant term.

Electricity and heat: The p-value for this variable is 0.000, that is less than 0.05. Therefore, the null hypothesis that the coefficient for this variable is equal to zero is rejected. This suggests that the electricity and heat variable have a significant effect on the dependent variable.

Building: The p-value for this variable is <0.001, that is less than 0.05. Therefore, the null hypothesis that this variable's coefficient is equal to zero is rejected. This suggests that the building variable has a significant impact on the dependent variable.

Transportation: The p-value for this variable is <0.001, that is less than 0.05. Therefore, the null hypothesis that this variable's coefficient is equal to zero is rejected. This suggests that the Transportation variable has a significant impact on the dependent variable.

Land use change and forestry: The p-value for this variable is 0.277 that exceeds 0.05. Therefore, the null hypothesis cannot be rejected that the coefficient value of this variable is zero. This suggests that the Land use change and forestry variable may not have a significant effect on the dependent variable. The omnibus test indicates that the fitted model with the independent variables electricity and heat, building, transportation, and land use change and forestry is a better fit than the intercept-only pattern. The value of chi-square likelihood ratio is 165.045 with 4 degrees of freedom and a significance level of .000 indicates strong evidence against the null hypothesis that the intercept-only model is a better fit, and in favor of the alternative hypothesis that the fitted model is a better fit.

5.2.2.2 Mean Impute Method

Table 8: Parameter Estimates

Parameter	B	Std. Error	95% Wald Confidence Interval		Hypothesis Test		
			Lower	Upper	Wald Chi-Square	df	Sig.
(Intercept)	1.446	7.8962	-14.030	16.923	.034	1	.855
(Intercept)	1.468	.1057	1.261	1.675	193.089	1	.000
electricity and heat	4.117	.9236	2.307	5.927	19.870	1	<.001
Building	.582	.1946	.201	.964	8.956	1	.003
Transportation	2.772	3.1642	-3.430	8.974	.767	1	.381
land use change and forestry	.001[a]	.0002	.000	.001			

The dependent variable, it is difficult to interpret the coefficients and the overall fit of the model. However, we can say that the intercept is not statistically significant as its p-value is greater than .05. The other hand, the p-values of the other independent variables is less than .05, which indicates that they are statistically significant in predicting the dependent variable. The Omnibus Test result shows a likelihood ratio chi-square of 161.676 with 4 degrees of freedom and a significant level of .000. This implies that the model as a whole is a good fit for the data and significantly explains variance in the dependent variable. Therefore, as a result, the model is statistically significant and thus rejects the null hypothesis.

5.2.2.3 KNN Method

Table 9: Parameter Estimates

Parameter	B	Std. Error	95% Wald Confidence Interval		Hypothesis Test		
			Lower	Upper	Wald Chi-Square	df	Sig.
(Intercept)	-22.987	16.8567	-56.025	10.052	1.860	1	.173
electricity and heat	1.055	.2304	.604	1.507	20.992	1	<.001
Building	5.778	1.9737	1.910	9.647	8.571	1	.003
Transportation	1.474	.4231	.645	2.304	12.143	1	<.001
land use change and forestry	10.421	6.7594	-2.828	23.669	2.377	1	.123
(Scale)	.003[a]	.0007	.002	.005			

This table shows the parameter estimates for a regression model. The intercept term has a coefficient of -22.987, which means that when all predictors are zero, the estimated value of the response variable is -22.987. The p-value for the intercept is not significant at the 0.05 level (p = 0.173), which suggests that the intercept may not be necessary in the model.

The coefficients for the predictor variables are as follows: electricity and heat (1.055), Building (5.778), Transportation (1.474), and land use change and forestry (10.421). These coefficients represent the changes in the expected value of the response variable for an increase of one unit in the corresponding predictor variable, holding every other predictor constant. All of these coefficients have significant p-values (less than 0.05, except Land use change and forestry), which suggests that they are important in predicting the variable of response.

Also included are the standard errors of the estimated coefficients which can be used to compute confidence intervals. The 95% confidence intervals for the predictor variables are also provided in the table. The test of Wald Chi-Square is used to test the hypothesis that each predictor variable has a coefficient of zero. The p-values for each test are also described in the table. The Omnibus Test results provided that a Likelihood Ratio Chi-Square value of 116.603, with 4 degrees of freedom, and a significance level of .000. This indicates that the fitted model (including all predictor variables) is a significantly better fit for the data than the intercept-only pattern. The p-value of .000 indicates that there is a very low probability of obtaining such a large test statistic under the null hypothesis that there is no difference between the models, supporting the alternative hypothesis that the fitted pattern is a better fit for the data.

5.2.2.4 PMM Method

Table 10: Parameter Estimates

Parameter	B	Std. Error	95% Wald Confidence Interval		Hypothesis Test		
			Lower	Upper	Wald Chi-Square	df	Sig.
(Intercept)	-11.458	9.3666	-29.816	6.900	1.496	1	.221
electricity and heat	1.251	.1267	1.002	1.499	97.420	1	.000
Building	4.987	1.0959	2.839	7.135	20.710	1	<.001
Transportation	1.054	.2332	.597	1.511	20.451	1	<.001
land use change and forestry	6.807	3.7545	-.552	14.166	3.287	1	.070
(Scale)	.001[a]	.0002	.001	.001			

The table shows the significant variables in this model are electricity and heat, building, and transportation, as their p-values are less than .05. The coefficient for electricity and heat is 1.251, indicating that for each unit increase in electricity and heat, the outcome variable is expected to increase by 1.251 units, on average, holding all other variables constant.

Similarly, the coefficients for building and transportation are 4.987 and 1.054, respectively. This indicates that for each unit increase in building and transportation, the outcome variable is expected to increase by 4.987 and 1.054 units, on average, holding all other variables constant. The intercept is not significant, indicating that it is not different from zero. The variable land use change and forestry is not significant at the .05 level. Overall, the pattern is significant with an omnibus test statistic of 151.419 and a p-value of 0.000. This indicates that the model explains a significance amount of the variability in the outcome variable.

5.2.2.5 EM Method

Table 11: Parameter Estimates

Parameter	B	Std. Error	95% Wald Confidence Interval		Hypothesis Test		
			Lower	Upper	Wald Chi-Square	df	Sig.
(Intercept)	-3.000	7.0670	-16.851	10.851	.180	1	.671
electricity and heat	1.393	.0949	1.207	1.579	215.500	1	.000
Building	4.415	.8266	2.795	6.035	28.527	1	<.001
Transportation	.745	.1748	.403	1.088	18.181	1	<.001
land use change and forestry	4.161	2.8322	-1.389	9.712	2.159	1	.142
(Scale)	.000[a]	.0001	.000	.001			

The table shows the significant variables in this model are electricity and heat, building, and transportation, as their p-values are less than .05. The intercept is not significant, indicating that it is not different from zero. The variable land use change and forestry is not significant at the .05 level. Overall, the model is significant with an omnibus test statistic of 168.272and a p-value of 0.000. This indicates that the model explains a significance amount of variability in the outcome variable.

5.2.3 Missing Data in the Bottom Data

5.2.3.1 Linear Trend at Point Method

Table 12: Parameters Estimates

Parameter	B	Std. Error	95% Wald Confidence Interval		Hypothesis Test		
			Lower	Upper	Wald Chi-Square	df	Sig.
(Intercept)	-5.849	7.1612	-19.884	8.187	.667	1	.414
electricity and heat	1.356	.0964	1.167	1.545	198.034	1	.000
Building	4.557	.8377	2.915	6.199	29.598	1	<.001
Transportation	.851	.1775	.503	1.199	22.988	1	<.001
land use change and forestry	4.805	2.8702	-.820	10.431	2.803	1	.094
(Scale)	.001[a]	.0001	.000	.001			

The table shows the significant variables in this model are electricity and heat, building, and transportation, as their p-values are less than .05. The intercept is not significant, indicating that it is not different from zero. The variable land use change and forestry is not significant at the .05 level. The pattern is significant with an omnibus test statistic of 168.272and a p-value of 0.000. This indicates that the model explains a significance amount of variability in the outcome variable.

5.2.3.2 Mean Impute Method

Table 13: Parameter Estimates

Parameter	B	Std. Error	95% Wald Confidence Interval		Hypothesis Test		
			Lower	Upper	Wald Chi-Square	df	Sig.
(Intercept)	-24.866	42.0680	-107.318	57.586	.349	1	.554
electricity and heat	.418	.5153	-.592	1.429	.659	1	.417
Building	11.266	5.0479	1.372	21.159	4.981	1	.026
Transportation	.421	.8728	-1.290	2.132	.233	1	.629
land use change and forestry	16.178	10.5701	-4.539	36.895	2.342	1	.126
(Scale)	266.043[a]	68.6921	160.388	441.298			

This table provides the results of gamma regression with identity link function after handling missing data with the linear trend at point method, row one of the table displays the intercept coefficient, which represents the response variable's value when all independent variables are equal to zero. Now, the intercept is statistically insignificant (p = 0.554), indicating that there is no evidence of a non-zero intercept. The next four rows show the estimated coefficients for each of the independent variables: electricity and heat, building, transportation, and land use change and forestry. The coefficients for electricity and heat, transportation and land use change are not significant (p >.05), implying that is no evidence of a linear relation between these variables and the response variable. The coefficient for building is statistically significant, implies that is a positive linear relation between this variable and the response variable. The model is significant with an omnibus test statistic of 71.282 and a p-value of 0.000. This indicates that the model explains a significance amount of variability in the outcome variable.

5.2.3.2 KNN Method

Table 14: Parameter Estimates

Parameter	B	Std. Error	95% Wald Confidence Interval		Hypothesis Test		
			Lower	Upper	Wald Chi-Square	df	Sig.
(Intercept)	-5.464	19.0747	-42.850	31.921	.082	1	.775
electricity and heat	1.081	.2592	.573	1.589	17.384	1	<.001
Building	5.839	2.2352	1.458	10.220	6.825	1	.009
Transportation	.784	.4696	-.136	1.704	2.788	1	.095
land use change and forestry	10.966	7.6418	-4.012	25.943	2.059	1	.151
(Scale)	.004[a]	.0009	.002	.006			

In this table, the electricity and heat and building variables have p-values lower than the significant level of .05, indicating that they are statistically significant predictors of the variable response. The transportation variable has a p-value of .095, which is slightly above the significant level. The land use change and forestry variable has a p-value of .151, which is not insignificant at the .05 level. The results of the omnibus test show that the likelihood ratio chi-square statistic is 106.405 with 4 degrees of freedom and a significant level of smaller than 0.05. This implies that the fitted model is significantly better than the intercept-only model, and therefore, this data fits the gamma regression model for the data after handling missing values.

5.2.3.4 PMM Method

Table 15: Parameter Estimates

Parameter	B	Std. Error	95% Wald Confidence Interval		Hypothesis Test		
			Lower	Upper	Wald Chi-Square	df	Sig.
(Intercept)	-5.615	11.1163	-27.402	16.173	.255	1	.613
electricity and heat	1.230	.1502	.936	1.525	67.054	1	<.001
Building	5.139	1.3012	2.589	7.689	15.598	1	<.001
Transportation	.820	.2747	.282	1.359	8.912	1	.003
land use change and forestry	7.651	4.4542	-1.079	16.381	2.950	1	.086
(Scale)	.001[a]	.0003	.001	.002			

The table shows the results of four explanatory variables: electricity and heat, building, transportation, and land use change and forestry. The model has a scale parameter of 0.001, which indicates a good fit. The intercept term is not significant, meaning that there is no constant effect on the emissions. The coefficients of electricity and heat, building, and transportation are non-negative and significance at the 0.01 level, illustrating that higher values of these variables are associated with higher emissions. The coefficient of land use change and forestry is positive but insignificant at the 0.05 level, suggesting that this variable has no effect on the emissions. The results of the omnibus test show that the likelihood ratio chi-square statistic is 139.921, with 4 degrees of freedom and a significant level of lower than 0.05. This suggesting that the fitted pattern is significantly better than the intercept-only model, and therefore, the model of gamma regression is a good fit for the data after handling missing values.

5.2.3.5 EM Method

Table 16: Parameter Estimates

Parameter	B	Std. Error	95% Wald Confidence Interval		Hypothesis Test		
			Lower	Upper	Wald Chi-Square	df	Sig.
(Intercept)	-5.894	7.0161	-19.646	7.857	.706	1	.401
electricity and heat	1.376	.0944	1.191	1.561	212.639	1	.000
Building	4.467	.8206	2.858	6.075	29.626	1	<.001
Transportation	.856	.1740	.515	1.197	24.203	1	<.001
land use change and forestry	4.354	2.8122	-1.158	9.866	2.397	1	.122
(Scale)	.000[a]	.0001	.000	.001			

The results indicate that electricity and heat, building, and transportation have significant positive effects on dependent variables, while land use change and forestry has a non-significant positive effect. On the other hand, the land use change and forestry are not significant. The scale parameter is estimated to be zero, indicating that there is no over dispersion in the model. The omnibus test for gamma regression is used to determine if the fitted model, the results show that the likelihood ratio chi-square statistic is 169.038, with 4 degrees of freedom and a significant level of less than 0.05. This represents the fitted model is significantly better than the intercept-only model, and therefore, the model of gamma regression is a good fit for the data after handling missing values.

5.3 Goodness of Fit

Now we will test the goodness fit method to determine the best method of handling missed data in gamma regression.

Table 17: Goodness of Fit

Missing Type	Method	Log Likelihood	AIC	BIC	Pseudo R
Missing in Top data	Linear Trend	-123.374	258.747	267.154	
	Mean. Impute	-116.607	245.214	253.621	0.9223
	KNN	-75.320	162.641	171.048	0.9969
	PMM	-77.183	166.366	174.774	0.996
	EM	**-73.319**	**158.638**	**167.045**	**0.9973**
Missing in Center data	Linear Trend	-81.831	175.662	-119.780	
	Mean. Impute	-81.831	175.662	184.069	0.9953
	KNN	-104.448	220.897	229.304	0.9790
	PMM	-86.895	185.791	194.198	0.9934
	EM	**-78.485**	**168.970**	**177.377**	**0.99625**
Missing in Bottom data	Linear Trend	-78.871	169.742	178.149	
	Mean. Impute	-126.323	264.646	273.053	0.9434
	KNN	-108.211	228.422	236.829	0.9706
	PMM	-92.045	196.090	204.497	0.99588
	EM	**-78.259**	**168.518**	**176.926**	**0.99634**

For Missing in the Beginning of Data: It's clear from the data that the mean impute method had the lowest log-likelihood value and the highest AIC and BIC values, indicating poor fit compared to other methods. KNN, PMM, and EM methods produced better fits with greatest log-likelihood values and a lower AIC and BIC values. Among these three methods, EM had the highest Pseudo-R^2 value (0.9973), indicating the best fit. The linear trend at point method also had a high Pseudo R^2value (0.988), indicating a good fit, but the log-likelihood, AIC, and BIC values were higher compared to KNN, PMM, and EM.

For the Missing Data Type "Missing in Middle Data," the results show that the EM imputation method has the highest log likelihood and AIC values, indicating the best fit, and the highest Pseudo R^2 value, indicating the best predictive power among the imputation methods for this missing data type. The linear trend, Mean Impute and PMM methods also performed well, while KNN performed relatively poorly. Overall, the Linear Trend and EM methods appear to be the most effective for imputing missing data in the middle of the dataset.

The K-Nearest Neighbor (KNN) Imputation Method: has a higher log-likelihood of -108.211, lower AIC of 228.422, and lower BIC of 236.829 than the mean imputation method. Its pseudo R^2 is also higher at 0.9706.

The Predictive Mean Matching (PMM) Method has the highest log-likelihood of -92.045 and the lowest AIC of 196.090 among all the methods. Its BIC of 204.497 is also lower than the KNN imputation method. Its pseudo R^2 is also relatively high at 0.99588.

The Expectation-Maximization (EM) Imputation Method has a log-likelihood of -78.259, AIC of 168.518, and BIC of 176.926, which are comparable to the linear trend method. Its pseudo R^2 is the highest among all methods at 0.99634. Overall, the results indicate that the Linear Trend and EM methods had the highest log likelihoods and Pseudo R^2 values, indicating better model fit. The AIC and BIC values were also lower for these methods, indicating a better balance between model fit and complexity.

Conclusion

From the practical study we noticed that

1- The EM and KNN imputation methods appear to be the best choices for handling missing values in the first quartile of the dataset, as they have higher log-likelihoods, lower AIC and BIC, and higher pseudo R^2 values than the other methods. However, the PMM method can also be a good choice due to its relatively high pseudo R^2 and lower computational cost compared to EM and KNN methods.

2- The EM imputation method appears to be the best choices for handling missing values in the central of the dataset, as it has a higher log-likelihoods, lower AIC and BIC, and higher pseudo R^2 values than the other methods.

3- The PMM and EM imputation methods appear to be the best choices for handling missing values in the bottom quartile of the dataset, as they have higher log-likelihoods, lower AIC and BIC, and higher pseudo R^2 values than the other methods. However, the KNN method can also be a good choice due to its relatively high pseudo R^2 and lower computational cost compared to PMM and EM methods.

4- Overall, the EM Method is the best fit method for the Gamma Regression model with missed data in any position (Beginning, Central, and the bottom of the data).

References

1- Abonazel, M. R., and Ibrahim, M. G. (2018)., "On estimation methods for binary logistic regression model with missing values"., International Journal of Mathematics and Computational Science, 4(3), 79-85.

2- Banerjee, P., & Seal, B. (2021)., "Partial Bayes Estimation of Two Parameter Gamma Distribution Under Non-Informative Prior", Statistics, Optimization & Information Computing, 10(4), 1110-1125. https://doi.org/10.19139/soic-2310-5070-1110.

3- Bossio, M., C., and Cuervo, E. C., (2015), "Gamma regression models with the Gammareg R package", Comunicaciones en Estadística, DOI: 10.15332/s2027-3355.2015.0002.05.

4- Cepeda-Cuervo, E., (2001). Modeling variability in generalized linear models, Departamento de Estadística.

5- Cepeda, C. E., Corrales, M., Cifuentes, M. V., & Zarate, H. (2016), "On gamma regression residuals", Journal of the Iranian Statistical Society, 15(1):29-44, DOI: 10.7508/jirss.2016.01.002.

6- Cox, D. R., & Reid, N. (1987), "Parameter Orthogonality and Approximate Conditional Inference",. Journal of the Royal Statistical Society. Series B (Methodological), 49(1), 1–39. http://www.jstor.org/stable/2345476

7- De Waal, T., Ooms, J. C., & Hart, J. (2011), "Predictive mean matching under model-based non-response", Journal of Official Statistics, 27(2), 277.

8- Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977), "Maximum Likelihood from Incomplete Data via the EM Algorithm", Journal of the Royal Statistical Society. Series B (Methodological), 39(1), 1–38. http://www.jstor.org/stable/2984875

9- Dupuy, J. F. (2020), "Censored Gamma Regression with Uncertain Censoring Status", Mathematical Methods of Statistics, 29(4), 172-196. DOI: 10.3103/s106653072004002x.

10- Keating, J.& Tripathy (2018), Gamma distribution, In Encyclopedia of statistical sciences (pp. 668-674). Wiley.

11- Kleinbaum, D. G., Kupper, L. L., & Muller, K. E. (2002), Applied regression analysis and other multivariable methods, Duxbury Press.

12- Little, R. J., & Rubin, D. B. (2019), Statistical analysis with missing data, John Wiley & Sons.

13- McLachlan, G., & Krishnan, T. (2008), The EM Algorithm and Extensions, John Wiley & Sons.

14- Meng, X. L., & Rubin, D. B. (1993)," Maximum likelihood estimation via the ECM algorithm: A general framework", Biometrika, 80(2), 267-278.

15- Rein DB, Wittenborn JS, Weinbaum CM, Sabin M, Smith BD, Lesesne SB., (2011), "Forecasting the morbidity and mortality associated with prevalent cases of pre-cirrhotic chronic hepatitis C in the United States", Dig Liver, 43(1):66-72. doi: 10.1016/j.dld.2010.05.006.

16- Robins, James M., Hernán, Miguel Ángel; Brumback, Babette,(2000), " Marginal Structural Models and Causal Inference in Epidemiology"., Epidemiology 11(5): pp 550-560.

17- Sigrist, F., & Stahel, W. (2011), "Using the Censored Gamma Distribution for Modeling Fractional Response Variables with an Application to Loss Given Default", ASTIN Bulletin: The Journal of the IAA, 41(2), 673-710. doi:10.2143/AST.41.2.2136992.

18- Van Buuren, S. (2012). Flexible Imputation of Missing Data. Chapman and Hall/CRC. https://doi.org/10.1201/b11826.

19- Zhipeng Sun, Guosun Zeng, and Chunling Ding. (2021), "Imputation for Missing Items in a Stream Data Based on Gamma Distribution. In Smart Computing and Communication), 5th International Conference, SmartCom, Paris, France,. Springer-Verlag, Berlin, Heidelberg, 236–247. https://doi.org/10.1007/978-3-030-74717-6_25.

تأثير معالجة البيانات المفقودة في متغير الاستجابة لانحدار جاما

أميرة إبراهيم الدسوقي

المستخلص:

البحث يقدم مقارنة شاملة لمختلف مناهج البيانات المفقودة في تحليل انحدار جاما. تقيم الدراسة أداء الاتجاه الخطي عند طريقة النقطة، وطريقة التضمين المتوسطة، وثلاث طرق احتساب متعددة في معالجة البيانات المفقودة في مواضع مختلفة (بداية ، وسط ، ونهاية) من نطاق البيانات. يتم استخدام تقنية تقدير الاحتمالية القصوى للتنبؤ بمعلمات نموذج انحدار جاما. تم تقديم مثال تجريبي لتوضيح تطبيق هذه الطرق في تحليل العوامل التي تؤثر على انبعاثات ثاني أكسيد الكربون في مصر. تكشف النتائج أن طرق التضمين المتعددة تتفوق في الأداء على الأساليب الأخرى من حيث الدقة.

الكلمات المفتاحية: انحدار جاما ، الاحتمالية القصوى ، البيانات المفقودة ، الاتجاه الخطي عند طريقة النقطة ، طريقة التضمين المتوسطة ، K- أقرب طريقة الجوار (KNN) ، مطابقة المتوسط التنبئي (PMM) ، حساب التعظيم من التوقعات (EM).